Protokoll

Evaluation von Genbank-Sequenzen

Sequenzähnlichkeit versus „Art“

Maximilian Vogel (1497058)
Email: maximilian.vogel@student.kit.edu

Benjamin Loritz (1459434)
Email: benjamin.loritz@student.kit.edu

Bearbeitungszeitraum: 16.02. –08.05.2015
Datum der Abgabe: 08.05.2015

Inhalt

1. Einleitung

1.1 Hintergrund

Schon seit je her war der Mensch bestrebt, Organismen und Pflanzen seiner Umwelt in Klassen einzuordnen und Familien, Gattungen und Arten zu identifizieren. Zur Zuordnung existieren viele Möglichkeiten – einige strukturierte Ansätze sind hierbei beispielsweise der morphologische Artbegriff nach Linné, der biologische oder populationsgenetische Artbegriff nach Mayr, das phylogenetische oder evolutionäre Artkonzept oder das chronologische Artkonzept, bei dem eine bekannte Klassifizierung um die Zeitkomponente erweitert wird.

Bei der verbreiteten, binären Nomenklatur (Linné) hängt die Klassifizierung von der Expertise und der Erfahrung einer Anzahl Taxonomen ab. Lohrmann und weitere dokumentieren (Lohrmann, 2012) eine immer weiter abnehmende Zahl von Experten und damit auch eine Abnahme entsprechender Publikationen. Diethard Tautz beschreibt diesen Fakt bereits 2003 (Diethard Tautz, 2003) als regelrechte „Krise der Taxonomie“, zumal auch die Erfahrung von 250 Jahren Forschung zwar gut dokumentiert, der Zugriff auf diese Daten sich jedoch schwierig und umständlich gestaltet. Ein Web-basierter Zugriff würde diese Schwierigkeit zwar abschwächen, dennoch haben einige Probleme nach wie vor Bestand. Eine grundsätzliche Schwäche des morphologischen Artkonzeptes ist, dass zwar Tiere und die meisten Pflanzen gut eingeordnet werden können, Bakterien und Protisten hingegen aufgrund der offensichtlich fehlenden oder schwer differenzierbaren äußeren Merkmale nicht klassifiziert werden können. David A. Caron et al beschreiben (David A. Caron, 2009) die Komplexität und Schwierigkeit der Einordnung von Protisten als zu schwierig und schlagen einen neuen Weg der taxonomischen Einordnung vor: angetrieben durch die stetige Verbesserung der DNA-Sequenzierung hinsichtlich Kosten, Geschwindigkeit und Qualität gilt es nun den wachsenden Datenbestand der bereits vorhandenen Sequenz-Informationen dazu zu nutzen ein besseres Verständnis der evolutionären Beziehungen zwischen verschiedenen Taxa zu gewinnen. Die Vorteile sind, dass sich DNA-Taxonomie auf ein großes Spektrum von Taxa, unabhängig vom jeweiligen Lebensabschnitt, einschließlich derer, die nur wenige morphologische Unterschiede aufweisen, anwenden lässt, sowie, dass die Zeit für die Ausbildung von Experten gespart werden kann, da ein standardisierter, automatisierter Ansatz zur Verarbeitung, Interpretation und Vergleich von Samples genutzt werden kann. Baxter identifiziert (Baxter, 2004) das größte Problem beim Zählen der Vielfalt von Taxa darin, dass die grundlegende Einteilung der Spezies nicht auf einer einzelnen Definition basieren und es viele unterschiedliche Konzepte gibt, die sich gegenseitig nicht vertragen oder zumindest in einigen Fällen scheitern. Beispiele hierfür sind das Scheitern des biologischen Artkonzepts bei allen asexuellen Abstammungslinien oder des Morphologischen bei reproduktiv isolierten Spezies, die sich aber kürzlich in Geschwistertaxa aufgeteilt haben.

1.2 Herangehensweise

Weit verbreitete Ansätze sind die Einordnung von Sequenz in Gruppen (Cluster), die entweder auf Ähnlichkeit der Sequenzen zu einer Referenzsequenz („Phylotyping“) oder auf Ähnlichkeiten innerhalb von Artengemeinschaften („*Operational Taxonomic Unit“* (*OTU*)) basieren. OTUs können als Gruppen von ähnlichen Sequenzen, die einen Repräsentanten besitzen, beschrieben werden. Es handelt sich dabei um einen Ansatz zur Bestimmung von Arten auf Sequenzebene, die jedoch nicht mit der binären Nomenklatur übereinstimmen müssen. Patrick D. Schloss und Sarah L. Westcott stellen (Westcott, 2011) mehrere Ansätze zur Interpretation und Implementierung von diesen so genannten OTU-basierten Systemen vor. Ein Vorteil hierbei ist, dass Daten sozusagen für sich selbst sprechen, da sie anhand ihrer Ähnlichkeiten untereinander in Cluster (OTUs) eingeteilt werden. Dies erschwert zwar einerseits die Namensgebung eines solchen Clusters, es gibt jedoch einige Beispiele, bei denen Organismen, die zur selben Art gehören, aber unterschiedliche Phänotypen aufweisen und umgekehrt Organismen mit gleichem Phänotyp zu unterschiedlichen taxonomischen Abstammungslinie zugeordnet werden müssen. Weitere Schwierigkeiten ergeben sich bei der Bestimmung eines Ähnlichkeitsgrenzwertes, also ab wie viel Abweichung in den Sequenzen die Sequenzen unterschiedlichen OTU’s zugeordnet werden müssen. Der vielleicht schwerwiegendste Nachteil entstehe jedoch bei der konkreten Anwendung von Algorithmen, da diese rechenintensiv und vergleichsweise langsam seien und viel Speicher benötigen. Die Wahl der Methode zur Gruppierung von Sequenzen sei ebenfalls problematisch – häufig benutzte Methoden in unterschiedlichen Disziplinen sind nächstes oder entferntestes, gewichtetes oder durchschnittliches hierarchisches Nachbarclustering.

Drive 5 stellt auf (Drive 5 - Bioinformatics Software and Services, 2015) einige bioinformatische Werkzeuge und Algorithmen zur Verfügung, die zur automatisierten Auswertung und Interpretation einer Menge von Daten genutzt werden können. *USEARCH* bietet beispielsweise die Möglichkeit des Clusterings auf einer Menge von Sequenzen mit zahlreichen Konfigurationsmöglichkeiten wie dem Ähnlichkeitsgrenzwert oder der Sortierung von Sequenzen. Die spannende Frage hierbei ist, welche Sequenzen wie aufbereitet werden müssen um sinnvoll verglichen werden zu können und wie die Ergebnisse interpretiert werden können. Des Weiteren bietet Drive 5 Algorithmen zum automatischen Alignment von Sequenzen (*Muscle*).

Eine große Hürde, die es in der DNA-Taxonomie zu überwinden gilt ist das Auffinden von vergleichbaren DNA-Sequenzen in den Genomen der zu vergleichenden Organismen. Tautz bringt in (Diethard Tautz, 2003) die Problematik zum Ausdruck: für unterschiedliche Zwecke eignen sich unterschiedliche Stellen im Genom unterschiedlich gut. Beispielsweise eignen sich die Gene, die für die kleine ribosomale Untereinheit kodieren am ehesten zur Unterscheidung von Gattungen, weniger

jedoch zur Unterscheidung von nahe verwandten Arten, da sie die größte, derzeit bekannte, taxonomische Abdeckung erzielen. Die mitochondriale Kontrollregion hingegen erlaubt hier eine schnelle Unterscheidung von verwandten Arten, eignet sich jedoch nicht zur Bestimmung von höheren taxonomischen Zugehörigkeiten. Eine gute Alternative könnten Regionen darstellen, die zwischen zwei konservierten Sequenzen liegen, die somit den Einsatz von universellen Primern erlauben. Doch in jedem Fall scheine es ratsam sich mehr als eine Sequenzregion anzuschauen um einen taxonomischen Status zuweisen zu können.

1.3 Das Projekt

Im vorliegenden Projekt finden sich nun viele der angesprochenen Problematiken wieder und sollen gelöst werden. Zunächst einmal gilt es ein sinnvolles, automatisiertes Verfahren zu finden, bei dem Datensätze auf Korrektheit hin überprüft werden und wie diese in eine Form gebracht werden können, sodass sie vergleichbar sind. Die weitaus spannendere Frage zielt jedoch auf die gefundenen Ergebnisse ab: Einerseits werden viele Sequenzen aus unterschiedlichen Regionen des Genoms untersucht und es stellt sich die Frage, in wie weit Ähnlichkeiten zwischen zwei spezifischen Arten, die bezüglich einer Region gefunden worden sind, sich auch in anderen Regionen wiederfinden und wie dies zu interpretieren ist, falls dies nicht der Fall ist. Gleiches gilt für die Unterschiede. Zusätzlich muss ein geeigneter Ähnlichkeitsgrenzwert gefunden werden und bestimmt werden, wie viele Unterschiede erlaubt sind, um in einem sinnvollen Taxonomie-Modell zu einer bestimmten Art gerechnet zu werden. Viele dieser Fragestellungen werden von der Betrachtungsweise und der Art der Aufbereitung der Daten abhängen und bedürfen einer eingehenden Untersuchung und Rückverfolgung zur Interpretation und der Manifestation von Ergebnissen.

Als Ergebnis soll einerseits der bestehende Artbegriff durch die auf DNA-Taxonomie beruhenden Auswertungen evaluiert werden und in diesem Zuge die Korrelation der auf Ähnlichkeit der Sequenzen basierenden Gruppen und der taxonomischen Gruppen, den so genannten Arten, festgestellt werden. Hierbei sollen problematische Gruppen identifiziert werden, für die die Identifizierung mit den entsprechenden Markern nicht möglich ist und für welche sogar das Artkonzept in Frage gestellt werden muss.

2. Material und Methoden

2.1 Der Workflow

In diesem Projekt gilt es ausgehend von einer Reihe von vergleichbaren Genbank-Sequenzen (bezüglich eines bestimmten Markers), eine Analysepipeline zu implementieren, an deren Ende Daten vorliegen, die eine Auswertung hinsichtlich Verwandtschaft der untersuchten Sequenzen zweier oder mehrerer Arten ermöglichen. Dabei werden sowohl statistische Ergebnisse, als auch Ergebnisse in Form von FASTA-Dateien erhoben, die in von der Pipeline abweichenden Schritten weiter verwendet oder zur genaueren Analyse zwischen den Zwischenschritten betrachtet werden können. Das zu implementierende Werkzeug bietet zusätzlich die Möglichkeit auch nur eine Teilmenge der Zwischenschritte auszuführen. Die Durchführung des Projekts orientiert sich im Wesentlichen am in Abbildung 2.1 gezeigten Workflows.

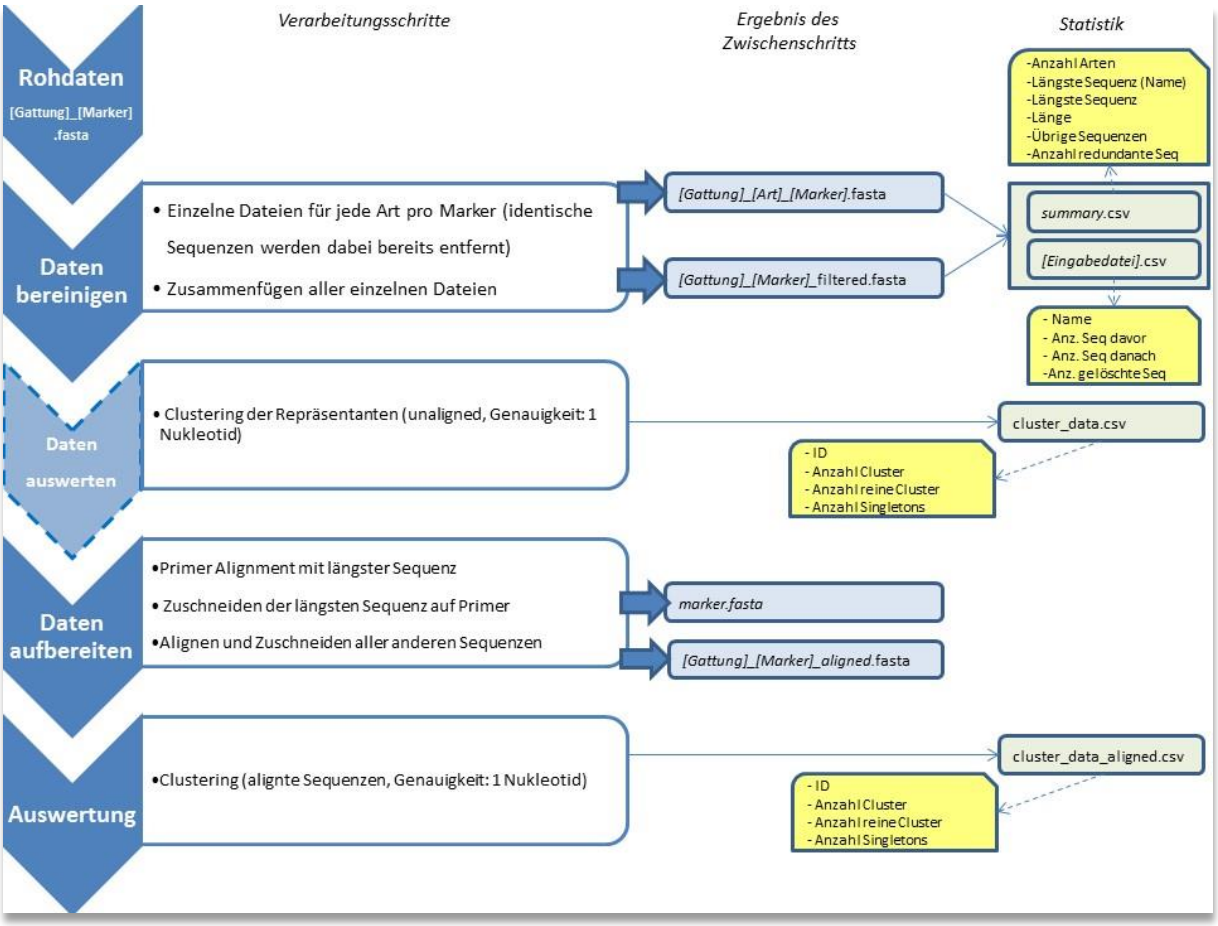


Abbildung 2.1 – Workflow der Analysepipeline

2.1.1 Rohdaten

Ausgangspunkt der Analysepipeline sind die Rohdaten. Innerhalb der Eingangsdatei müssen die Sequenzen im FASTA-Format vorliegen. Eine Zeile, die mit „>“ beginnt, enthält den Namen der Sequenz sowie eine Beschreibung der Daten: `>Name der Sequenz|ID|Marker|`. Alle weiteren Zeilen bis zur nächsten, die mit „>“ beginnt, enthalten dann die eigentliche Nukleotidsequenz.

2.1.2 Daten bereinigen

Im zweiten Schritt müssen die Daten bereinigt werden. Das bedeutet im Einzelnen, dass identische Sequenzen innerhalb einer Art eliminiert werden müssen, um den Datenbestand auszudünnen ohne Information zu verlieren. Dazu werden zunächst alle gleichlautenden Artnamen-Sequenzen in einzelne Dateien gruppiert, sodass diese nun nur noch ausschließlich alle zu einer Art gehörenden Sequenzen enthält. Dieser Schritt des Auftrennens ist essenziell, da man beim Eliminieren von identischen Sequenzen andernfalls Gefahr laufen würde, identische Sequenzen zweier unterschiedlicher Arten zu löschen. Dies würde einen wichtigen Informationsverlust bedeuten. Zusätzlich werden nun alle Sequenzen gelöscht, die vollständig in einer anderen Sequenz der Art-Datei enthalten sind. Es werden also nur die längsten Sequenzen behalten, die sich nicht gegenseitig enthalten. Dies entspricht dem 100%-Clustering-Modus von USEARCH, bei dem ein einzelnes, unterschiedliches Nukleotid zweier Sequenzen dazu führt, dass mindestens zwei Cluster entstehen. Die gefundenen Sequenzen sind Repräsentanten für die in ihr enthaltenen Sequenzabfolge und enthalten mindestens dieselbe Information, weshalb die enthaltenen Sequenzen gelöscht werden können ohne Information zu verlieren.

➔ **Ergebnis**:

1. Eine Datei pro Art mit Dateinamen gemäß folgendem Schema: *[Gattung]_[Art]_[Marker].fasta*
2. Eine Datei, die alle eindeutigen Sequenzen einer Gattung enthält (es können auch mehrere Repräsentanten-Sequenzen für eine Art enthalten sein!) gemäß folgendem Schema: *[Gattung]_[Marker]_filtered.fasta*

➔ **Statistik**:

1. *summary.csv*: In dieser Datei sind die folgenden Werte des Bereinigungsvorgangs hinterlegt:
    - Anzahl Arten
    - Längste Sequenz (Name)
    - Längste Sequenz (Nukleotidabfolge)

- Länge der längsten Sequenz
- Übrige Sequenzen
- Anzahl identischer Sequenzen unter den übrigen Sequenzen

2. *[Eingabedatei].csv*: In dieser Datei sind folgende Werte des Bereinigungsvorgangs hinterlegt:
    - Name
    - Anzahl Sequenzen davor
    - Anzahl Sequenzen danach
    - Anzahl gelöschte Sequenzen (entspricht #Seq davor - #Seq danach)

2.1.3 Daten auswerten

Im dritten Schritt werden die bereinigten Daten ausgewertet. Dazu wird mittels USEARCH ein Clustering der Repräsentanten einer Gattung bezüglich eines bestimmten Markers durchgeführt. Der entscheidende Punkt an dieser Stelle ist die Einstellung der Schrittweite des Ähnlichkeitsgrenzwertes von USEARCH. Diese wird mittels der Formel

$$Schrittweite = \frac{1}{Länge\ der\ Sequenz}$$

ermittelt und somit ein Clustering auf 1 Nukleotid Genauigkeit erreicht. Es handelt sich hierbei um ein Zwischenergebnis, bei dem beispielsweise die unterschiedlichen Längen der Sequenzen noch nicht berücksichtigt sind – die Daten sind also nicht aligned.

➔ **Statistik:**

1. *cluster_data.csv*: In dieser Datei sind die folgenden Werte der Zwischenauswertung (für jede Gattung/Marker) hinterlegt:
    - ID
    - Anzahl Cluster
    - Anzahl reine
    - Länge der längsten Sequenz

2.1.4 Daten aufbereiten

Im vierten Schritt werden die Daten aufbereitet und auf eine vergleichbare Form gebracht. Zunächst wird versucht die zu einem Marker gehörenden Primer an die längste Sequenz zu alignen

und die überstehenden Enden (vor dem forward-Primer und nach dem reverse-Primer) zu beschneiden oder, wenn der Primer über das Ende hinaus geht, mit „N“ aufzufüllen. Hier muss man mit großer Sorgfalt herangehen: Die frei verfügbaren Programme bieten keine speziellen Modi für das Alignment von Primern, sodass hier ein selbst geschriebenes Programm zum Einsatz kommt. Dieses alignt die Primer naiv indem es bspw. den forward-Primer in die Sequenz hineinschiebt und die Anzahl übereinstimmender Nukleotide zählt und das Verhältnis zur Primerlänge zurückgibt. Dies wird bis zur Mitte der Sequenz durchgeführt und anschließend die beste Position zum Zuschneiden verwendet. Dies schließt in der aktuellen Implementierung auch Fälle ein, in denen nur wenige Positionen übereinstimmen.

Die zugeschnittene Sequenz wird schließlich dazu verwendet, sie mit jeder anderen Sequenz aus dem Datensatz zu alignen, was in unserem Fall mit ClustalO geschieht. Für das erneute Clustering werden entstandene Gaps wieder entfernt.

2.1.5 Auswertung

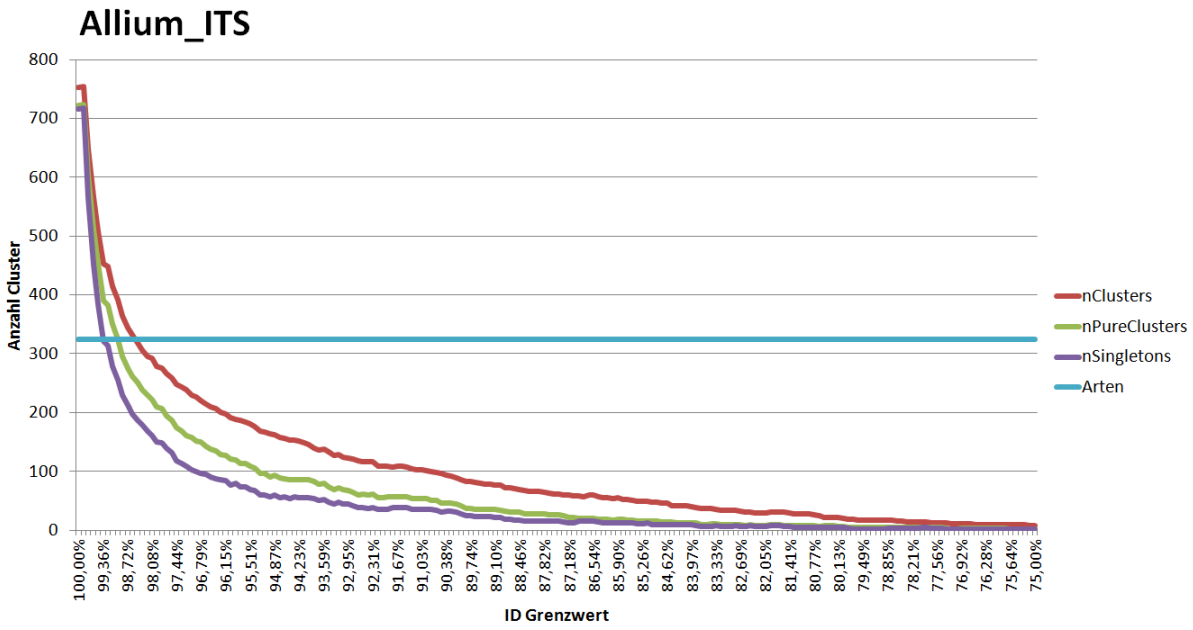
Mit dem nun bereinigten Datensatz wird erneut mit USEARCH geclustert, wie in 3. beschrieben. Die Idee dabei ist, dass die Sequenzen nun nur noch relevante Nukleotide enthalten, die sich tatsächlich zwischen den Primern befinden. Die Primer sind dabei in der zugeschnittenen Sequenz noch vorhanden.

3. Ergebnisse

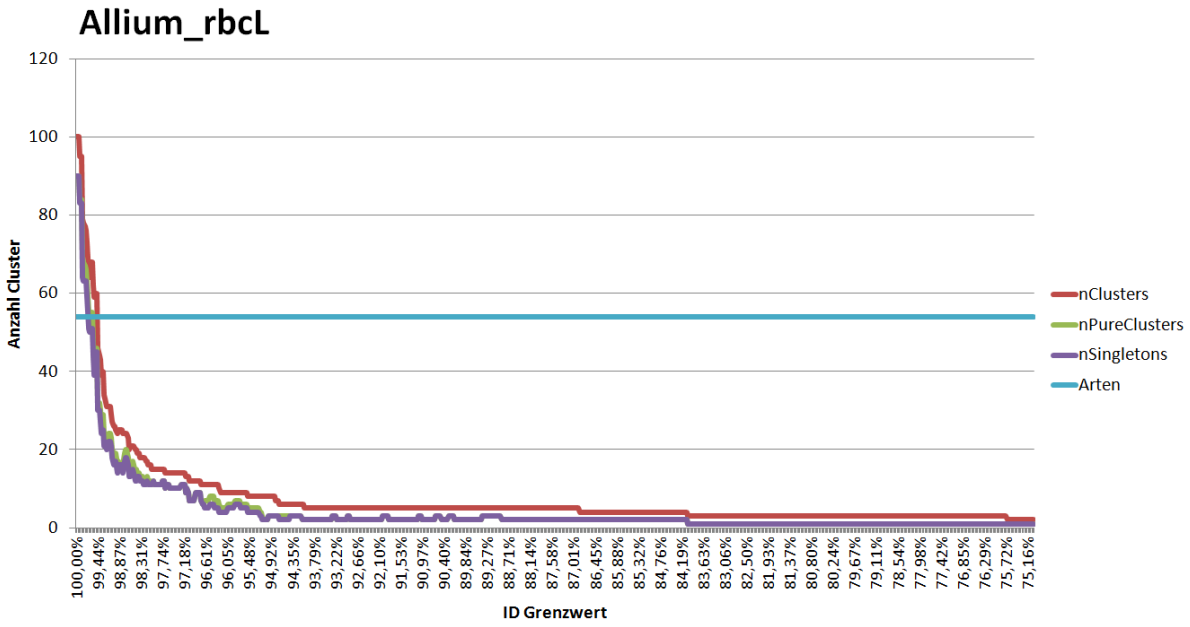
In diesem Projekt wurden beispielhaft Sequenzen der Gattungen *Allium* und *Lycium* ausgewertet. Es liegen zu Beginn die Datenbanksequenzen mehrerer Arten beider Gattungen bezüglich dreier Marker (ITS, rbcL, psbA_trnH) vor. Die folgenden Auswertungen und Ergebnisse stammen von den Schritten der Kapitel 2.1.3 „Daten auswerten“ und 2.1.5 „Auswertung“ des in Kapitel 2.1 vorgestellten Workflows und beziehen sich im Fall von 2.1.3 auf die Auswertung der bereinigten Daten und im Fall von 2.1.5 auf die Endergebnisse der Analysepipeline.

3.1 Ergebnisse zu Kapitel 2.1.3: „Daten auswerten“

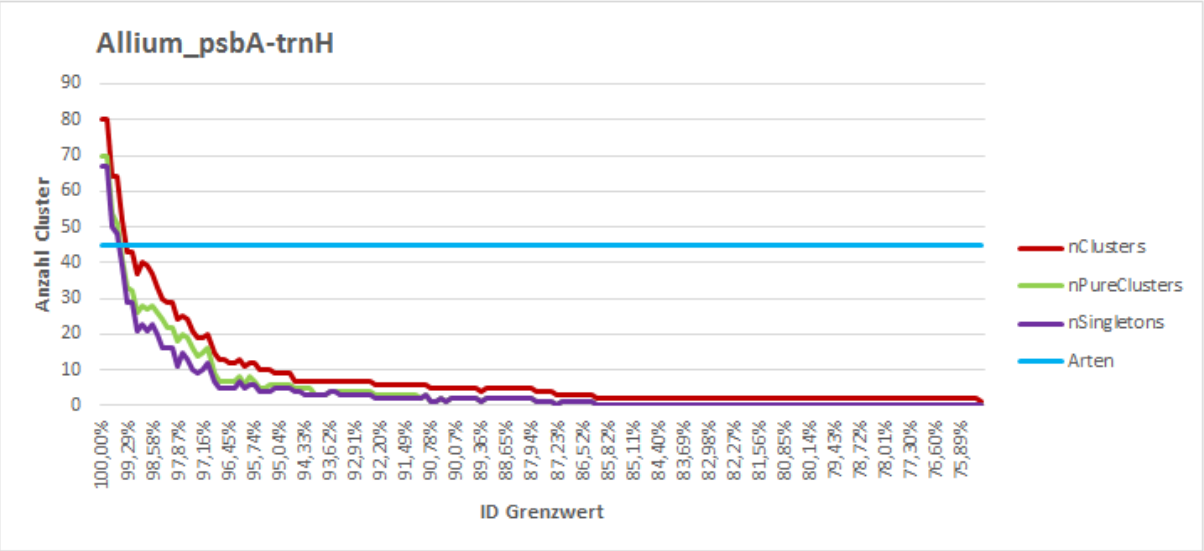
In den folgenden Grafiken sind vier Kurven aufgetragen – eine Konstante markiert die Anzahl der Arten, die übrigen drei Kurven zeigen die Anzahl der Cluster in Abhängigkeit vom Ähnlichkeitsgrenzwert. Dabei wird unterschieden, wie viele Cluster es insgesamt gibt („nCluster“), wie viele dieser Cluster ausschließlich Sequenzen einer Art enthalten („nPureCluster“) und wie viele der *nCluster* einelementig sind („Singletons“), also nur eine einzelne Sequenz enthalten.



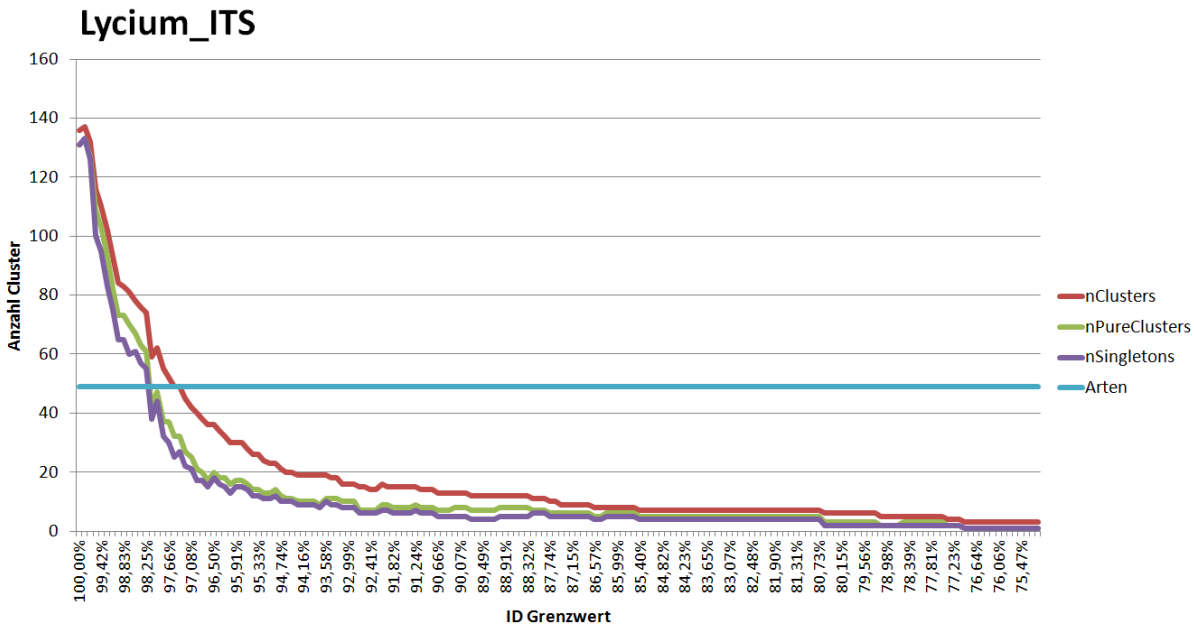
Zusammenfassung:	
Anzahl Sequenzen:	921
Anzahl Arten:	325
Längste Sequenz (Name):	Allium_thunbergii|KM051462.1|ITS
Länge längste Sequenz:	780
Übrige Sequenzen:	815 (88,5%)
Anzahl redundanter Sequenzen in übrigen Sequenzen:	44



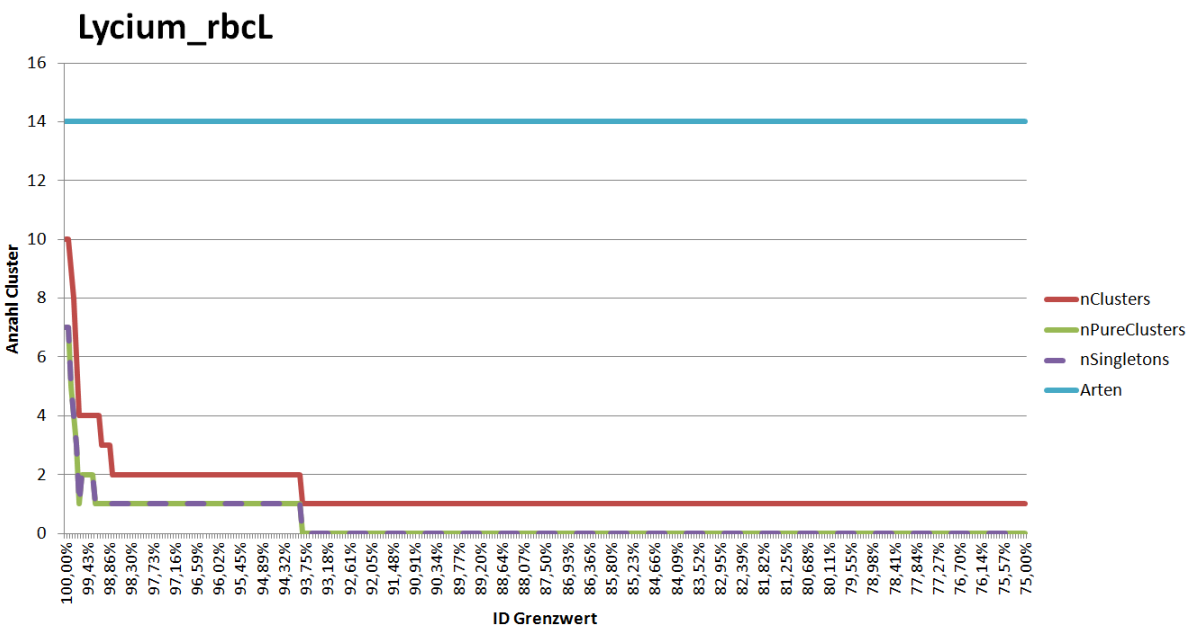
Zusammenfassung:	
Anzahl Sequenzen:	195
Anzahl Arten:	54
Längste Sequenz (Name):	Allium_fistulosum|AB292285.1|rbcL
Länge längste Sequenz:	3188
Übrige Sequenzen:	121 (62%)
Anzahl redundanter Sequenzen in übrigen Sequenzen:	25



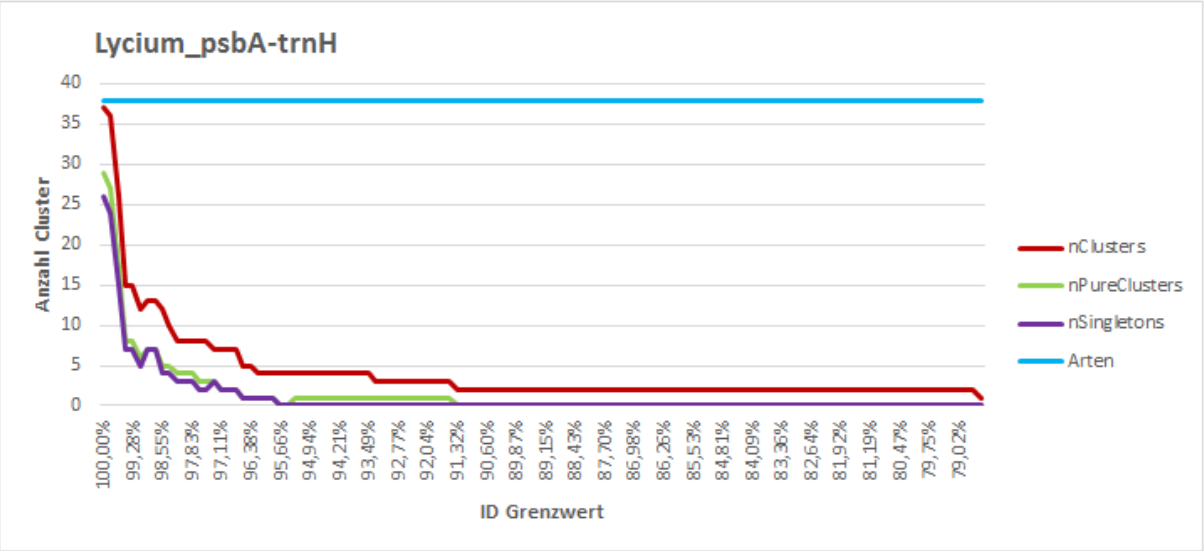
Zusammenfassung:	
Anzahl Sequenzen:	153
Anzahl Arten:	45
Längste Sequenz (Name):	Allium_monanthum|KC704218.1|psbA-trnH
Länge längste Sequenz:	705
Übrige Sequenzen:	109 (71%)
Anzahl redundanter Sequenzen in übrigen Sequenzen:	17



Zusammenfassung:	
Anzahl Sequenzen:	153
Anzahl Arten:	49
Längste Sequenz (Name):	Lycium_barbarum|JQ320141.1|ITS
Länge längste Sequenz:	685
Übrige Sequenzen:	141 (92,1%)
Anzahl redundanter Sequenzen in übrigen Sequenzen:	7

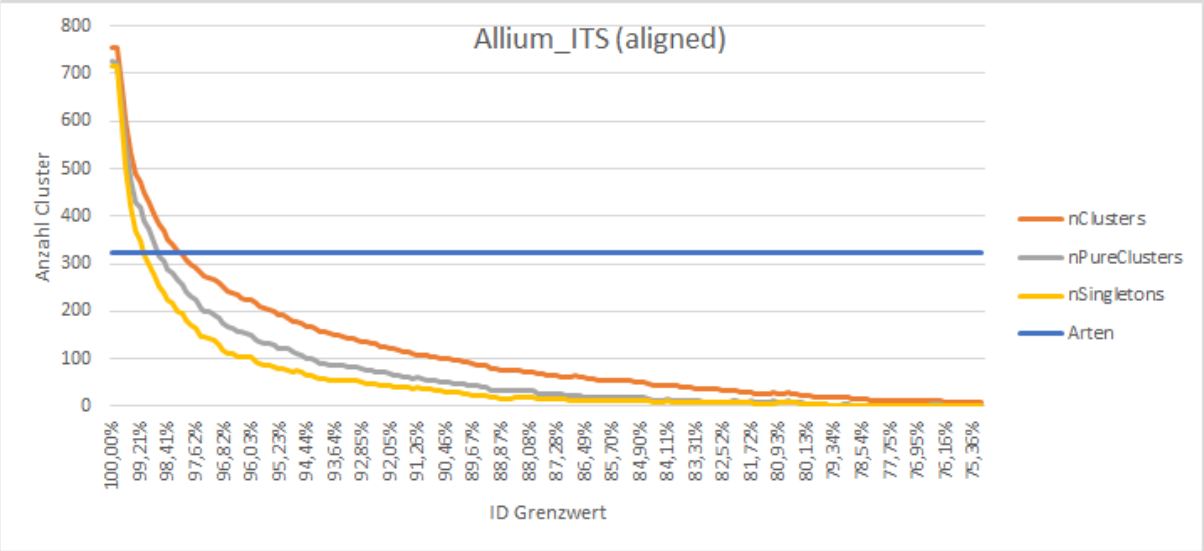


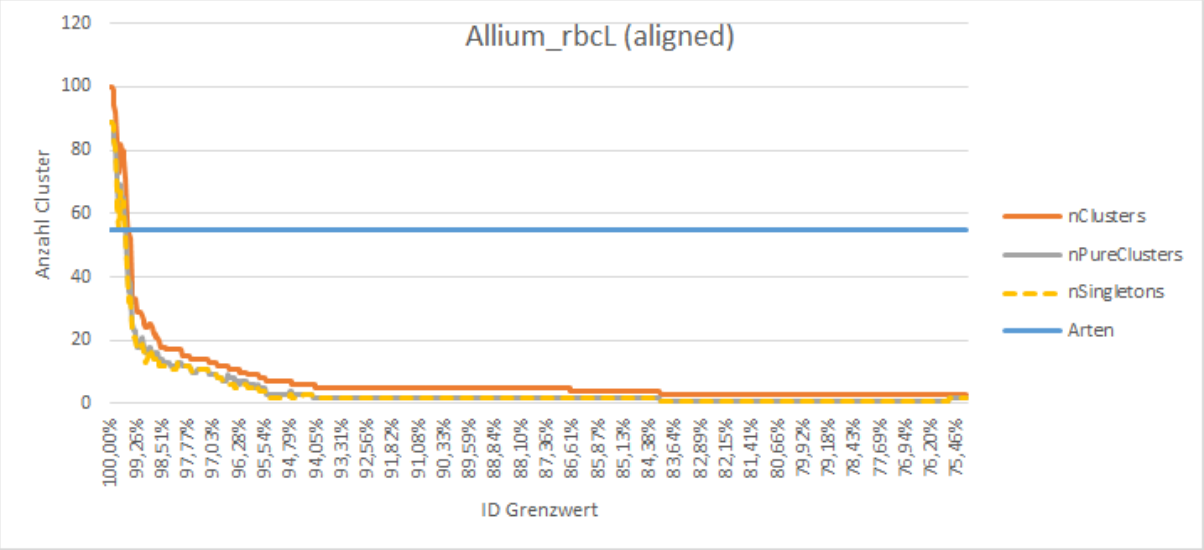
Zusammenfassung:	
Anzahl Sequenzen:	25
Anzahl Arten:	14
Längste Sequenz (Name):	Lycium_cestroides|U08613.1|rbcL
Länge längste Sequenz:	1408
Übrige Sequenzen:	17 (68%)
Anzahl redundanter Sequenzen in übrigen Sequenzen:	5

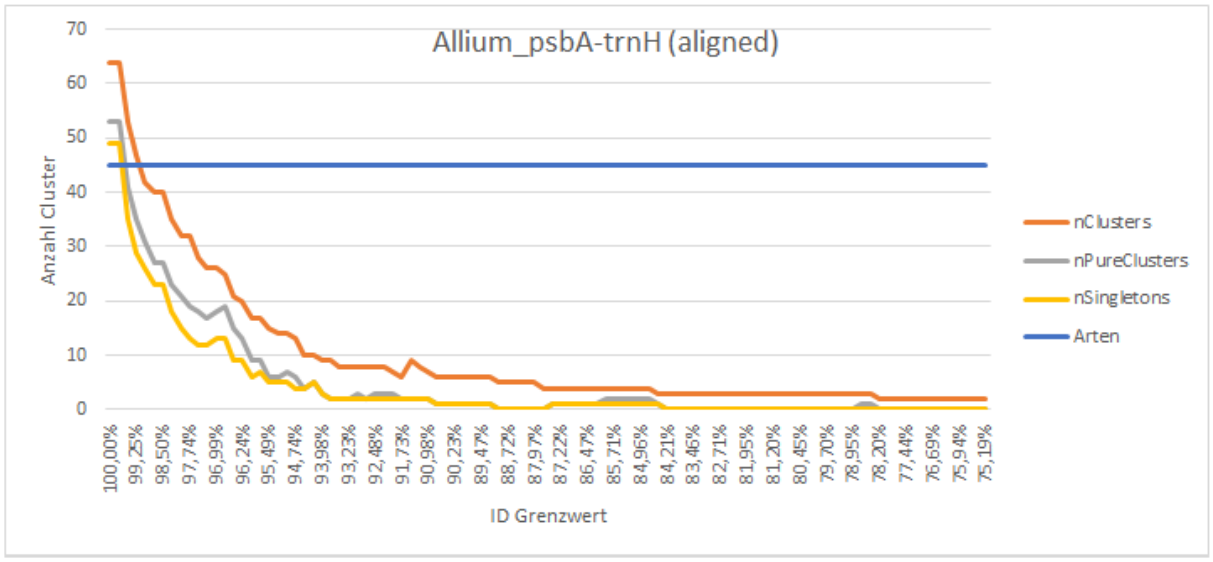


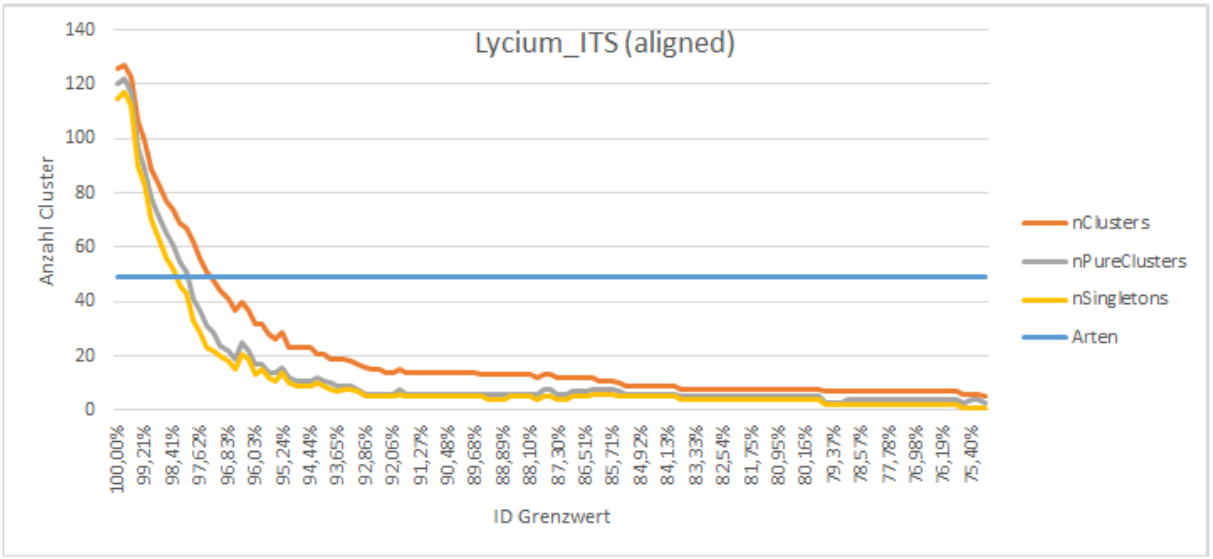
Zusammenfassung:	
Anzahl Sequenzen:	66
Anzahl Arten:	38
Längste Sequenz (Name):	Lycium_shawii|HM195013.1|psbA-trnH
Länge längste Sequenz:	553
Übrige Sequenzen:	60 (91%)
Anzahl redundanter Sequenzen in übrigen Sequenzen:	11

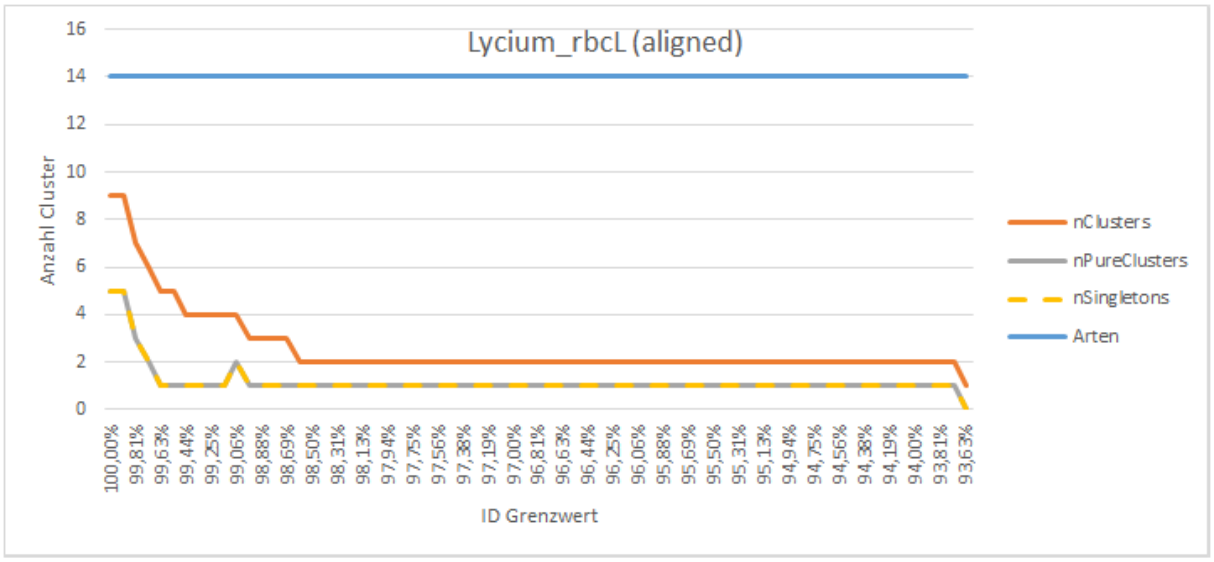
3.2 Ergebnisse zu Kapitel 2.1.5: „Auswertung“

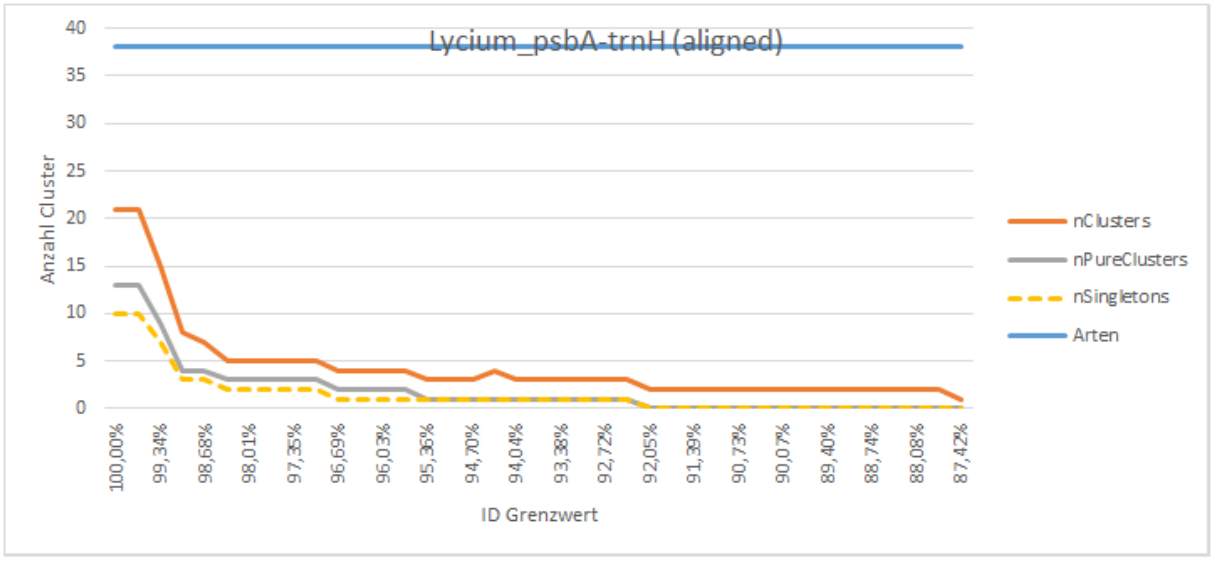





Allium_psbA-trnH (aligned)
Anzahl Cluster
70
60
50
40
30
20
10
0
100,00%
99,25%
98,50%
97,74%
96,99%
96,24%
95,49%
94,74%
93,98%
93,23%
92,48%
91,73%
90,98%
90,23%
89,47%
88,72%
87,97%
87,22%
86,47%
85,71%
84,96%
84,21%
83,46%
82,71%
81,95%
81,20%
80,45%
79,70%
78,95%
78,20%
77,44%
76,69%
75,94%
75,19%
ID Grenzwert
nClusters
nPureClusters
nSingletons
Arten


Lycium_ITS (aligned)
Anzahl Cluster
140
120
100
80
60
40
20
0
100,00%
99,21%
98,41%
97,62%
96,83%
96,03%
95,24%
94,44%
93,65%
92,86%
92,06%
91,27%
90,48%
89,68%
88,89%
88,10%
87,30%
86,51%
85,71%
84,92%
84,13%
83,33%
82,54%
81,75%
80,95%
80,16%
79,37%
78,57%
77,78%
76,98%
76,19%
75,40%
ID Grenzwert
nClusters
nPureClusters
nSingletons
Arten


Lycium_rbcL (aligned)
Anzahl Cluster
ID Grenzwert
nClusters
nPureClusters
nSingletons
Arten


Lycium_psbA-trnH (aligned)
Anzahl Cluster
ID Grenzwert
nClusters
nPureClusters
nSingletons
Arten

4. Diskussion

4.1 Diskussion zu den Ergebnissen aus 2.1.3: „Daten auswerten“

Bei der Auswertung der Ergebnisse zu 2.1.3.lässt sich allgemein sagen, dass einerseits die Anzahl der Cluster mit abnehmendem Ähnlichkeitsgrenzwert auch abnimmt. Dies ist nicht weiter verwunderlich, da eine Menge von Sequenzen auf eine abnehmende Anzahl Cluster verteilt wird, wenn man Sequenzunterschiede schrittweise zulässt. Des Weiteren fallen die Kurven für reine Cluster und Singletons mit abnehmendem Ähnlichkeitsgrenzwert zusammen, da die Anzahl der Singletons, also diejenigen Cluster, die nur eine einzelne Sequenz enthalten, geringer wird und gleichzeitig die Anzahl der reinen Cluster (in denen die Singletons ja enthalten sind) ebenfalls weniger werden (aber immer mehr Sequenzen enthalten). Sie werden gewissermaßen „verunreinigt“ je größer sie werden.

Generell fällt bei mehreren Auswertungen (alle bis auf Allium ITS (Auflösung zu gering)) auf, dass die Kurve nicht monoton fällt. Während die Schwankungen bei Allium rbcL, Lycium rbcL und Lycium psbA-trnH nur bei reinen und Singleton Cluster auftreten, steigt bei Allium psbA-trnH und Lycium ITS zwischendurch auch die Anzahl der Cluster insgesamt wieder an. Worauf dies zurückzuführen ist, kann hier nicht geklärt werden. Die Annahme, dass die Reihenfolge, in welcher die Sequenzen in USEARCH verarbeitet werden, dabei eine Rolle spielt, liegt nahe. Man könnte die Sequenzen in jedem Durchgang derart umsortieren, dass Sequenzen, die in einen Cluster fallen, untereinander stehen.

Bei Lycium mit rbcL-Marker ergibt die Auswertung, dass die Kurve der reinen Cluster und die der Singletons identisch ist, was bedeutet, dass Cluster mit mehr als einer Sequenz keine reine Cluster sind und den Schluss nahe legen, dass es weniger Arten gibt als bisher angenommen. Wie auch im folgenden Absatz, muss man allerdings die vorhandene Anzahl an Sequenzen für diesen Marker in Betracht ziehen und abwägen, ob eine solche Aussage zulässig ist. Um zu sehen, ob man ähnliche Beobachtungen auch bei anderen Gattungen und Markern machen kann, ließe sich die Anzahl der Sequenzen und Gattungen und Markern künstlich auf eine ähnliche Menge reduzieren und den Workflow mit dem reduzierten Datensatz wiederholen.

Des Weiteren fällt bei Lycium mit rbcL wie auch psbA-trnH auf, dass es selbst bei einem Ähnlichkeitsgrenzwert von 1 weniger Cluster gibt als Arten für diesen Marker in der Datenbank vorhanden sind. Dies kann mehrere Ursachen haben:

- Manche Pflanzen wurden falsch bestimmt.
- Es gibt weniger Arten als bisher angenommen.

Aufgrund der insgesamt geringen Anzahl an Sequenzen (25 gegenüber 921 bei Allium ITS oder 195 bei Allium rbcL) sollt man hier jedoch keinen Schluss ziehen.

Ein ähnlicher Aspekt sind die Schnittpunkte der einzelnen Kurven mit der konstanten Gerade der Anzahl Arten. Je höher der Genauigkeitsgrenzwert beim Schnittpunkt, desto wahrscheinlicher ist es, dass es tatsächlich so viele Arten gibt. Allerdings entsteht dann die Frage, von welcher Kurve man den Schnittpunkt betrachtet und wie man die Werte der anderen Kurven an dieser Stelle interpretiert.

4.2 Diskussion zu den Ergebnissen aus 2.1.4: „Daten aufbereiten“

Primeralignment: Das in Schritt 3 beschriebene naive Primeralignment funktioniert in der Praxis nicht gut genug, mit Ausnahme von Allium ITS sind die Übereinstimmungen im Bereich 40% bis 60% und die Position mit dem jeweils besten Wert liegen selten am „Rand“. Um beim Primeralignment aussagekräftigere Ergebnisse zu bekommen, könnte man an folgenden Punkten ansetzen:

a. Bei Überlappungen über den Rand hinaus müssen mindestens 50% des Primers in der Sequenz liegen.
b. Ausgeben mehrerer Alignmentkandidaten, unter anderem mit Anzahl der übereinstimmenden Nukleotiden aber vor allem auch der längsten zusammenhängenden Folge von Nukleotiden. Die Anzahl der übereinstimmenden Nukleotide sollte man im Verhältnis zur Anzahl der überlappenden Nukleotide angeben und nicht im Verhältnis zur Länge des Primers.

Der Schritt des Primeralignments lässt sich nur mit guten Daten vollends automatisieren, was in der Praxis nicht der Fall sein wird bzw. bei den verwendeten Daten nicht der Fall ist. Hier wäre ein interaktiver Modus des Programms nützlich, welcher dem Benutzer die Möglichkeit gibt, zwischen verschiedenen Alignmentkandidaten auszuwählen oder aber, falls keiner der Kandidaten gut genug ist, die nächste Sequenz auszuprobieren.
Des Weiteren lässt sich an dieser Stelle des Workflows auch ein manuell zugeschnittenes Alignment verwenden, da sich die finale Auswertung im nachfolgenden Schritt separat ausführen lässt.

Histogramm: An der Stelle im Workflow, an dem die Sequenzen alle aligned sind, gibt es die Möglichkeit, die einzelnen Positionen über alle Sequenzen hinweg auszuwerten, in dem man beispielsweise ein Histogramm über die Nukleotide und Gaps erstellt. Dadurch könnte man einen weiteren Einblick in die Daten erhalten und nachvollziehen, durch welche Positionen oder Regionen unterschiedliche Cluster entstehen.

4.3 Diskussion zu den Ergebnissen aus 2.1.5: „Auswertung“

Vergleicht man die Grafiken von den Rohdaten mit den alignten Daten, so fallen bei Allium ITS und Allium rbcL wenig Unterschiede auf, wohingegen bei Allium psbA-trnH einige Unterschiede festzuhalten sind: Zum einen sind die Anzahl der Cluster zu Anfang niedriger und zum anderen gibt es im Gegensatz zu den Rohdaten zu Beginn mehr reine als Singleton Cluster. Zur Erinnerung: Nur bei Allium ITS ist das Primeralignment akzeptabel. Es ist daher gerade bei Allium psbA-trnH möglich, dass durch das Zuschneiden zu viel von den Sequenzen weggeschnitten wurde.

Bei den Plots von Lycium ITS lassen sich keine nennenswerte Veränderungen zwischen den Roh-und den alignten Daten beobachten. Bei Lycium rbcL hingegen fällt auf, dass bereits bei 93% nur noch ein Cluster existiert, was auch bei den zugehörigen Rohdaten selbst bei 75% noch nicht erreicht wurde. Dies liegt vermutlich daran, dass durch das Zuschneiden mit schlechten Primeralignments Information verloren gegangen ist. Bei Lycium psbA-trnH können ähnliche Beobachtungen gemacht werden wie bei Lycium rbcL und Allium psbA-trnH: Es sind von Anfang an weniger Cluster und es wird bereits vor der 75%-Marke ein einzelnes Cluster erreicht, was bei den anderen nicht der Fall war.

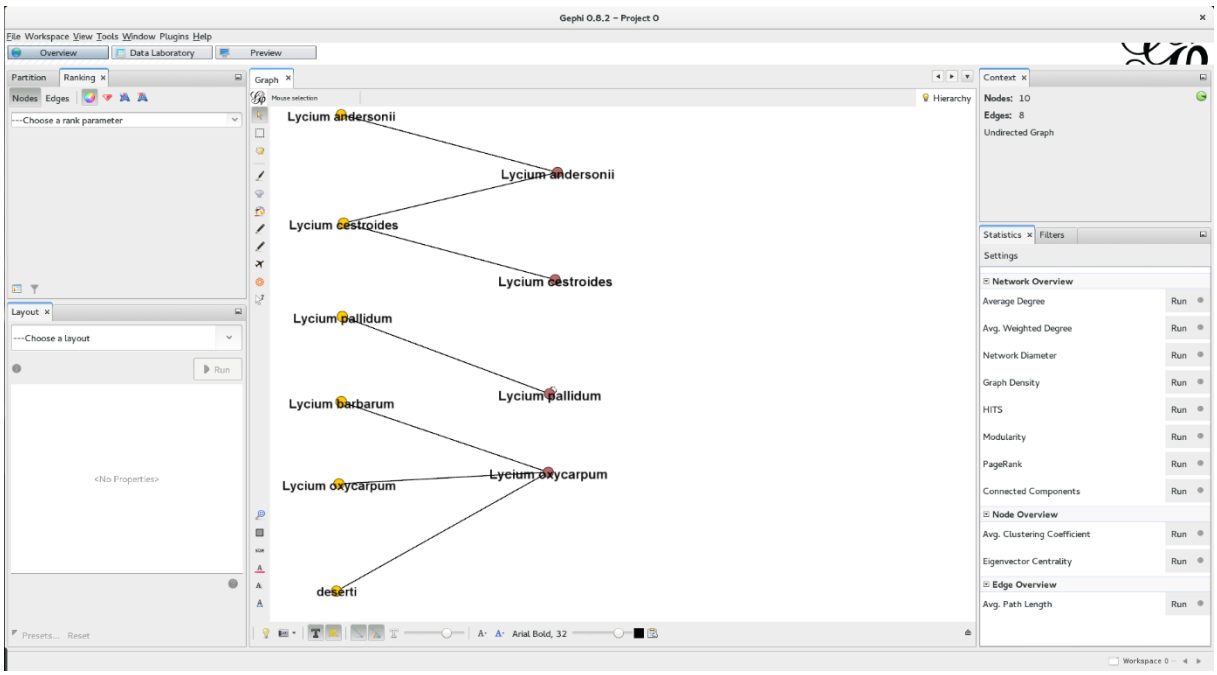
4.4 Diskussion / Ausblick

Statistische Auswertung: Eine sinnvolle Erweiterung der vorliegenden Implementierung wäre in Kapitel 2.1.2 eine zusammenfassende Statistik zu erzeugen, die die Rohdaten anhand von Merkmalen wie beispielsweise der Anzahl der Sequenzen, den Mittelwert der Sequenzlängen oder den GC-Gehalt beschreibt. Anschließend könnten dann die Veränderungen zwischen den Rohdaten und den bereinigten Daten durch eine sinnvolle Darstellung veranschaulicht werden.

Modellierung als Netzwerk: Aus welchen Sequenzen sich die einzelnen Cluster je nach Genauigkeitsgrenzwert zusammensetzen, wird im aktuellen Workflow nicht tiefer analysiert. Ist man daran interessiert, so sollte man das Clustering mit dem relevanten Abhängigkeitswert wiederholen und dabei die –uc-Option von USEARCH aktivieren. Die resultierende Datei ließe sich dann manuell untersuchen. Dies ist als ein Umstand und vor allem nicht gerade nutzerfreundlich anzusehen. Ein kleines selbstgeschriebenes Programm (parse_uc) kann bereits die angesprochenen Dateien von USEARCH einlesen und verarbeiten. Aktuell werden dadurch die Anzahl der reinen Cluster bestimmt. Während es zum einen möglich ist, an dieser Stelle auch die Artnamen der reinen Cluster oder aber auch verteilter Arten auszugeben, könnte man zum anderen die Beziehung über die einzelnen Clusterdurchgänge hinweg als Netzwerk modellieren. Ein Punkt bzw. Knoten stellt einen Cluster mit einem festen Genauigkeitsgrenzwert dar. Zwischen zwei Clustern mit unterschiedlichen Genauigkeitsgrenzwerten existiert genau dann eine Verbindung (oder auch Kante), wenn mindestens eine Sequenz aus dem einen Cluster auch in dem anderen Cluster enthalten ist. Umgekehrt ließen sich zusätzlich auch Arten als Punkt modellieren und eine Verbindung zu einem Cluster ziehen, wenn diese

Art im Cluster enthalten ist. Bei der Modellierung als Netzwerk sind hier also viele Möglichkeiten denkbar und man müsste sich genau überlegen, welche Zusammenhänge bzw. Relationen man modellieren möchte. Die Modellierung als Netzwerk bietet dabei folgende Möglichkeiten zur weiteren Analyse: Zum einen ließen sich Netzwerkanalysetools verwenden (networkx, NetworKit, graph-tool, …) und zum anderen könnte man die so entstandenen Netzwerke mit grafischen Programmen wie zum Beispiel Gephi visualisieren. Der folgende Screenshot ist nur ein Mockup, soll aber veranschaulichen, wie die Graphenvisualisierung in Gephi eingesetzt werden könnte, um die Ergebnisse einfacher zu interpretieren. (Dies bedingt, dass man die Ergebnisse in Netzwerkmodell überführt und in ein Format bringt, welches Gephi einlesen kann.)

Korrelation über mehrere Marker: Eine weitere interessante Analyse welche man mit den Ergebnissen des Clusterings durchführen könnte, ist die Korrelation der Cluster zwischen verschiedenen Markern zu berechnen. Hierbei muss man allerdings beachten, dass man dann die Schrittweite des Ähnlichkeitsgrenzwertes nicht mehr in Abhängigkeit der Sequenzlänge wählt, sondern für alle Marker einheitlich. Entstehen die gleichen Cluster bei unterschiedlichen Markern oder korrelieren zumindest „gut“, so wären Aussagen und Interpretationen über die Anzahl Arten fundierter. Die Korrelation der Cluster würde dabei über die enthaltenen Sequenzen bestimmt werden und was „gut“ in diesem Kontext bedeutet, müsste man untersuchen. Dabei sollte man auch beachten, dass man ungefähr die gleiche Menge Sequenzen für jeden Marker besitzt, im Idealfall sogar die gleichen Akzessionen.



5. Literaturverzeichnis

Baxter, M. L. (2004). *The promise of a DNA taxonomy.* Edinburgh: The Royal Society.

David A. Caron, P. D. (September 2009). Defining DNA-Based Operational Taxonomic Units for Microbial-Eukaryote Ecology. *APPLIED AND ENVIRONMENTAL MICROBIOLOGY*, S. 5797–5808.

Diethard Tautz, P. A. (Februar 2003). A plea for DNA taxonomy. *TRENDS in Ecology and Evolution*, S. 70ff.

Drive 5 - Bioinformatics Software and Services. (04. April 2015). Von http://www.drive5.com: http://www.drive5.com abgerufen

Lohrmann, V. (29. Mai 2012). Taxonomische Forschung in Deutschland - Eine Übersichtsstudie. Universität Potsdam.

Westcott, P. D. (May 2011). Assessing and Improving Methods Used in Operational Taxonomic Unit-Based Approaches for 16S rRNA Gene Sequence Analysis. *APPLIED AND ENVIRONMENTAL MICROBIOLOGY*, S. 3219–3226.